

MRMCsamplesize: An R Package for Estimating Sample Sizes for Multi-Reader Multi-Case Studies

Dennis Robert^{*1}, Saigopal Sathyamurthy¹, Preetham Putha¹

¹Qure.ai, India

^{*}Corresponding author

Abstract

Multi-Reader Multi-Case (MRMC) studies are typically used to evaluate improvement in diagnostic accuracy of readers (diagnosticians) when they are assisted by a computer-assisted device (CAD) such as, but not limited to, those based on Artificial Intelligence (AI) algorithms. Statistical analysis of MRMC study data is not trivial and these studies can consume a lot of resources. Optimal planning is crucial and estimation of sample size is a significant step during the study planning phase. MRMC sample size estimations require many parameter assumptions and without pilot data this is generally not intuitive. MRMCsamplesize package can help researchers to estimate sample sizes for an MRMC study in the absence of any pilot data. The program outputs the number of cases required for a given number of readers. The package can also estimate sample sizes for scenarios where intra-cluster correlation (ICC) needs to be adjusted.

1. Introduction

Computer-assisted detection (CAD) software devices in healthcare are meant to assist clinicians or diagnosticians (here onwards referred to as 'readers') in improving their diagnostic accuracy. Artificial Intelligence/Machine Learning (AI/ML) based CAD devices in the field of Medical AI are becoming increasingly popular. Within the field of Medical AI, the most significant breakthroughs are occurring in the field of Medical Imaging because of a number of reasons, such as the availability of digitized medical images, the image being a representation of the anatomical area of interest in a structured form without much ambiguity as opposed to electronic medical record data and the availability of deep neural networks specifically useful for image pattern recognition such as Convolutional Neural Networks (CNNs). As of August 2023, roughly 500 AI/ML-enabled medical devices have been granted marketing authorization by FDA and approximately 75% of them are intended to be used in the field of Radiology (FDA 2023). A reasonable take on how AI/ML will impact the field of radiology is well captured in the quote, "Radiologists who use AI will replace those who don't", and not that AI will replace radiologists (Langlotz 2019). Multi-Reader Multi-Case (MRMC) studies are very important in this context because these studies are conducted to test the hypotheses of whether using CAD tools such as those based on AI/ML can improve the diagnostic accuracy of readers. These studies are also accepted by regulatory bodies such as FDA as a means for clinical performance assessment of CAD devices for pre-market notification. (CDRH 2022). MRMC studies are expensive to conduct and consume a lot of resources. The statistical analysis is not trivial due to complex correlation structures. Estimation of a reasonable sample size (both the number of readers and the number of cases) is a crucial element in the study planning phase and this requires making reasonable assumptions of a number of different parameters.

To our knowledge, no R packages are currently available for the estimation of sample sizes for a planned MRMC study without prior pilot data. For example, **RJafroc** (Chakraborty and Zhai 2023) is a comprehensive package to facilitate various types of reader study analysis and it also includes functions to estimate sample sizes. However, it requires pilot data to facilitate input argument assignments, and in the vast majority of scenarios, researchers will have no access to any such pilot data. Even though we can still use functions in **RJafroc** to work without pilot data, this requires the users to supply variance components which are very difficult to conjecture. **MRMCAov** (Smith, Hillis, and Pesce 2023; Smith and Hillis 2020) and **iMRMC** (Gallas 2023) are both useful for the statistical analysis of MRMC study data, but they do not contain functions for the estimation of sample sizes to aid study planning. There are JAVA-based software programs available for sizing an MRMC study. One example is an open-source JAVA-based graphical user interface (GUI) program (Hillis and Schartz 2018) developed at the University of Iowa. This program allows users to perform sample size estimation for MRMC studies and it can do so for a wide variety of options such as with and without

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

pilot data, various types of study designs such as random readers and random cases, fixed readers and fixed cases, etc., different methods such as both OR and DBM methods. Another JAVA-based GUI program named **iMRMC** (Gallas, Pennello, and Myers 2007), developed by the same research team who developed the **iMRMC** package, is also available for sample size estimations. Without pilot data, it might not be intuitive to supply variance components or correlation components which are required as inputs for sample size estimations in these programs. In addition, intra-cluster correlation (ICC) due to the presence of multiple target lesions within a diseased case (for example, the presence of



Figure 1: A hypothetical example of an MRMC study workflow with a single reader and five images. The order in which the reader reads the images is typically randomized in each session to mitigate reading-order bias. *Washout period of 4 weeks is typically used in cross-over designs, but sequential designs without any washout period may also be used depending on the intended use of the CAD/AI device. It is also not rare to mix reading sessions with cases from both modalities. For example, Reading Session 1 can include a mix of images with and without the assistance of AI and similarly for Reading Session 2 as well. In such scenarios, if an image is read by a reader with AI assistance in the first session, then in the second session, that image has to be read by that reader without AI assistance.

multiple lung nodules in a chest X-ray image) further complicates the analysis because the diagnostic accuracy is dependent not just on detecting a lesion *somewhere* in a case, but also localizing it correctly. In such scenarios, sample size estimations have to be adjusted for the anticipated ICC in diseased cases. Throughout this article we use the term “diseased cases” for cases with target lesion of interest and “non-diseased cases” for cases without any target lesion of interest.

MRMCsamplesize (Robert 2023), the R package of focus in this article, can estimate the number of required cases for a given number of readers for a fully-crossed MRMC study using the Obuchowski-Rockette (OR) method (Nancy A. Obuchowski and Rockette 1995). It can be used for sizing MRMC studies with or without adjustment for anticipated ICC. The sample size output from **MRMCsamplesize** was validated by comparing it against published literature. The results of this validation exercise are also presented in this article.

2. Fundamentals of MRMC Study Design and Statistical Analysis

In an MRMC study, a set of readers interpret a set of cases (typically medical images) with and without AI. A hypothetical example of a fully-crossed MRMC study workflow with a single reader and five images is illustrated in Figure 1.

Since the same reader reads the same image twice and the same images are read by different readers, there are intra- and inter-reader variabilities which result in complex correlation that needs to be considered during statistical analysis. The choice of whether the design has a sequential nature or cross-over nature can also have an impact on sample size estimations (Nancy A. Obuchowski and Bullen 2022).

While multiple different statistical methods are proposed for the analysis of MRMC study data, two methods are particularly popular. They are:

- Obuchowski-Rockette (OR) method
- Dorfman-Berbaum-Metz (DBM) method

The OR method fits a correlated-by-error-test-by-reader ANOVA (a mixed effects ANOVA model), treating readers as random effects, the interaction between test and reader as another random effect and a fixed effect for the test as independent variables and average per-test reader performance outcomes such as area under the receiver operating characteristics curve (AUC) as the dependent variable. (Nancy A. Obuchowski and Rockette 1995). The DBM method fits a test-by-reader-by-case conventional ANOVA to case-specific pseudovalues (Dorfman, Berbaum, and Metz 1992). OR method is considered to be more intuitive since the parameters are more interpretable due to the method modelling the observed reader-performance outcomes rather than the pseudovalues (Iowa 2023). **MRMCsamplesize** package is developed based on the OR method.

3. Sample Size Estimation for MRMC Studies

3.1. Endpoints

The primary endpoints for MRMC studies are typically the difference in reader performance as quantified by a figure-of-merit (FOM) such as AUC, sensitivity (Se), specificity (Sp) or area under the

free-response curve (FROC) paradigm such alternative-FROC (AFROC).

MRMCsampleize can estimate sample size if the planned primary endpoint is difference in AUC or Se. The methodology of sample size estimation is detailed in the next section.

3.2. Detailed Methodology

It has to be known during the study planning whether the diseased cases can contain multiple target lesions of interest or not. For example, multiple lung nodules can be present in a single chest X-ray image and the reader's accuracies are evaluated by the number of correct lesions that the reader identifies. In this case, there can be correlations between the lesions in the same diseased case, and the resulting ICC has to be thus adjusted. The first step is to estimate a required number of diseased cases (cases with target lesions of interest) and non-diseased cases (cases with no target lesions of interest) assuming independence (i.e., no adjustment for ICC) and then adjust the numbers by accounting for the design effect (DE) resulting from the anticipated ICC (Nancy A. Obuchowski and Hillis 2011).

3.2.1. Assuming independence between observations The methods described here are largely based on the methods published by Obuchowski, Rockette and Hillis in multiple publications (Nancy A. Obuchowski and Rockette 1995; Nancy A. Obuchowski 2000; Hillis and Schartz 2018; Zhou, Obuchowski, and McClish 2011). The words 'case' and 'image' are used interchangeably in this article.

Assuming that the design is fully-crossed and that there are two modalities to be studied (for example, each reader reads each case twice, one with AI and one without AI), the null and alternative hypothesis is:

$$H_0 = \mu_1 = \mu_2 \quad H_A = \mu_1 \neq \mu_2 \quad (1)$$

In (1), μ_i is the mean FOM of readers for the modality i .

For the sample size estimation, under the alternative hypothesis, it is specified that $\mu_1 - \mu_2 = \Delta$, where Δ is the suspected difference (effect size) in the FOM of interest. Obuchowski and Rockette proposed a modified F statistic to test the null hypothesis using a 2-way mixed effects ANOVA model (Nancy A. Obuchowski and Rockette 1995). The noncentrality parameter of the F-distribution, denoted by λ , can be used to derive sample size estimates with pre-specified power and type 1 error rate to be able to detect a pre-specified minimum effect size of Δ . λ is given by:

$$\lambda = \frac{J\Delta^2}{2 \left(\sigma_b^2 (1 - r_b) + \frac{\sigma_w^2}{K} + \sigma_c^2 [(1 - r_1) + (J - 1)(r_2 - r_3)] \right)} \quad (2)$$

The definitions of the parameters in the RHS (Right-Hand Side) of equation (2) are listed in Table 1.

The estimated power p of a study with J readers is:

$$p = \text{Prob} \left(F_{t-1, (t-1)(J-1); \lambda} > F_{1-\alpha; t-1, (t-1)(J-1)} \right) \quad (3)$$

Where $F_{t-1, (t-1)(J-1); \lambda}$ denotes a random variable having a non-central F distribution with degrees of freedom $t - 1$ and $(t - 1)(J - 1)$ and non-centrality parameter of λ , t is the number of modalities used in the study and $F_{1-\alpha; t-1, (t-1)(J-1)}$ is the $(1 - \alpha)$ 100th percentile of a central F distribution with the same degrees of freedom.

The non-centrality parameter of F distribution, λ , that would provide a pre-specified power p and type I error rate α can be calculated. This calculation needs numerator and denominator degrees of freedom also to be pre-specified in addition to required p (typically 0.8) and α (typically 0.05). The numerator and denominator degrees of freedom used for calculating λ are respectively $t - 1$ and $(t - 1)(J - 1)$ (Zhou, Obuchowski, and McClish 2011). For example, λ that would provide 80% power with a type 1 error rate of 5% are 18.12, 12.36, 9.92 and 8.72 when J is equal to 4, 6, 10 and 20, respectively.

σ_b^2 and σ_w^2 are difficult to conjecture in the absence of pilot data but are very important parameters that determine the power of the study. Conjecturing the range of FOMs is much more intuitive. If the FOMs of the readers are assumed to follow a normal distribution, the relationship between range and standard deviation can be used to derive the variances σ_b^2 and σ_w^2 by multiplying the conjectured range by a constant derived from normal distribution (Nancy A. Obuchowski 2000). Another alternative approach is to use the simple rule that standard deviation equals the range divided by four as approximately 95% of the values are distributed within four standard deviations. The former method is more conservative, especially when the number of readers is less than 30-35, and thus may be the more safer approach. This is the default method used in **MRMCsampleize** to derive inter- and

Table 1: Parameters and definitions

Parameter	Definition	Comments
J	Number of readers.	A minimum of 5 is recommended for any MRMC study.
Δ	The difference in FOM (effect size) of readers when using one modality as compared to the other modality.	In sample size estimations, Δ has a great effect on sample size. Typically used Δ range from 0.04 to 0.06 corresponding to a 4 to 6 percent difference in the FOM of readers between the two modalities. Smaller the Δ , larger the required number of cases and readers.
σ_b^2	Variance of the FOM of readers when using the same modality for the same sample of cases (inter-reader variability).	This is not intuitive to conjecture as this is an estimate of variance.
σ_w^2	Variance of the FOM of a reader when using the same imaging technique for the same sample of cases in different occasions (intra-reader variability).	This is not intuitive to conjecture as this is an estimate of variance.
σ_c^2	Variance of the FOM between different sample of cases (case sample variability) and is a function of the number of cases in the sample.	
r_1	Correlation between FOMs of readers when same sample of cases are evaluated by the same reader using different modalities.	Based on a systematic review of about 32 studies examining 49 different comparisons by Rockette et. al. (Rockette et. al. 1999), the range of r_1 was found to be from 0.35 to 0.59. The average value was 0.47 and this is used by Obuchowski to derive the reference sample tables in a reference paper (Obuchowski 2000). Lower values are conservative.
r_2	Correlation between FOMs when the same cases are evaluated by different readers using the same modality.	For sample size estimations, it is recommended to consider $r_2 = r_3$. In such a scenario, only r_1 and r_b needs to be conjectured for sample size estimations among the four correlation parameters.
r_3	Correlation between FOMs when the same cases are evaluated by different readers using different modalities.	For fully-crossed study design, typically: $r_1 \geq r_2 \geq r_3 \geq 0, r_b \geq 0$
r_b	The correlation between FOMs when the same readers evaluate cases using different modalities	A value of 0.8 is typically used as per recommendation by Rockette. et. al (Rockette et. al. 1999)
K	Number of times each reader interprets each case using the same modality.	Often, $K = 1$ because each reader evaluates a case only once using one modality.

intra-reader variances from the intra- and inter-reader variability ranges conjectured.

The only remaining unknown parameter in equation (2) is σ_c^2 and this can be computed by re-arranging equation 3. σ_c^2 is in turn, a function of the anticipated average FOM (AUC or Se). In the case when the FOM of interest is AUC, σ_c^2 is given by:

$$\sigma_c^2 = \frac{\left(0.0099 \times e^{-\frac{A^2}{2}}\right) \times \left(5A^2 + 8 + \frac{A^2+8}{R}\right)}{N_D} \quad (4)$$

where A is:

$$A = \phi^{-1}(\theta) \times 1.414 \quad (5)$$

In equation (5), θ is the anticipated average AUC of readers and $\phi^{-1}(\theta)$ is the inverse cumulative normal distribution function. R is the ratio of non-diseased cases to diseased cases and N_D is the number of diseased cases. Note that equations (4) and (5) are based on estimating variance of AUC assuming a binormal distribution as described by Zhou, Obuchowski and McClish (Zhou, Obuchowski, and McClish 2011). Blume (Blume 2009) recommended another approach to estimate the variance of AUC which requires no parametric assumptions. This is also implemented in the [MRMCsamplesize](#) package.

When the FOM is sensitivity (Se), σ_c^2 can be estimated by considering the properties of a binomial proportion estimate and its variance as given by:

$$\sigma_c^2 = Se \times (1 - Se) / N_D \quad (6)$$

From here onwards, the equations will be based on the assumption that the FOM of interest is AUC. For Se, the only difference will be re-arrange the equations (2) and (6). Using (2) and (4), the number of diseased cases required for the MRMC study can be estimated by:

$$N_D = \left(\frac{\left(0.0099 \times e^{-\frac{A^2}{2}}\right) \times \left(5A^2 + 8 + \frac{A^2+8}{R}\right)}{\frac{I\Delta^2}{2\lambda} - \left(\sigma_b^2(1 - r_b) + \frac{\sigma_w^2}{K}\right)} \right) \times ((1 - r_1) + (J - 1)(r_2 - r_3)) \quad (7)$$

Note that in (7), the term $\sigma_b^2(1 - r_b)$ is also known as the *test-by-reader-variance* in some literature and it is the variance of the interaction between the modality and the reader (Nancy A. Obuchowski and Hillis 2011).

The total sample size (sum of diseased and non-diseased cases), N_T required for the planned study would thus be:

$$N_T = N_D \times (1 + R) \quad (8)$$

3.2.1.1. Estimating inter- and intra-reader variances using range One can employ two methods for deriving σ_b^2 and σ_w^2 using corresponding ranges, $range_b$ and $range_w$, respectively, the latter two being much more intuitive to conjecture in the absence of any pilot data.

range_b: The anticipated difference between the highest accurate (highest FOM) reader in the study and the lowest accurate (lowest FOM) reader.

range_w: The anticipated difference between the FOMs of a reader who interprets the same cases using the same modality at two different times.

The first method assumes that readers' FOMs follow a normal distribution and thus σ_b and σ_w can be computed by multiplying $range_b$ and $range_w$ by constants c_b and c_w derived from the normal distribution. Note that c_b depends on the number of readers, but c_w is dependent on the number of modalities used in the study (typically, there are two modalities or reading sessions per reader). Three statistical properties are used for the estimation of c_b and c_w :

- The expected range for a sample of size n in a symmetric distribution with mean 0 is twice the expected largest value in a sample of the same size.
- The density of the largest value X_n in a sample of size n from a distribution with density f and cumulative distribution function F is $n \times f(x) \times F(x)^{n-1}$
- For a normal distribution sample with size n , the expected range is the expected range of a standard normal sample of size n times the standard deviation.

So the expected largest value X_n in a sample of size n following a normal distribution is obtainable by integration as per order statistics.

$$E(X_{(n)}) = \int_{-\infty}^{\infty} n \times x \times f(x) \times F(x)^{n-1} dx \quad (9)$$

$$c_b = \frac{1}{2E(X_n)} \quad (10)$$

As an example, when J is 10 and $range_b$ is assumed to be 0.1 (10% difference in FOM between the highest and lowest accurate readers), σ_b is calculated by:

$$c_b = 0.3249$$

$$\text{and } \sigma_b = range_b \times c_b = 0.03249$$

c_w , as discussed above, is dependent only on the number of reader sessions per reader and this is typically always 2 in a fully crossed MRMC study assessing accuracies of two modalities. Hence c_w is 0.8862 in most scenarios regardless of J . The calculation of $sigma_w$ from c_w is trivial.

The second method is rather trivial and uses the relationship between range and standard deviation.

$$\sigma_b = \frac{range_b}{4} \quad (11)$$

$$\sigma_b = \frac{range_b}{4} \quad (12)$$

Note that in order to get the variances σ_b^2 and σ_w^2 , the standard deviations σ_b and σ_w must be squared.

Between the two methods, first method is the more conservative method especially when number of readers (J) is less than 30.

3.2.2. Adjusting for intra-cluster correlation For taking into account the fact that there can be multiple target lesions in a single case and it is required to evaluate the accuracy of readers based on detecting all the lesions in all the diseased cases, one must conjecture an anticipated ICC and the average number of target lesions in diseased cases (s). The design effect (DE) is a function of ICC and s . This will yield the N_{Dc} which is the number of diseased cases after adjustment for the ICC . Note that N_{Dc} is always less than or equal to N_D and this means that number of diseased cases required for a MRMC study where ICC is expected is less than the number that is required where this adjustment is not required given the same assumptions for all other parameters. An ICC of 0.5 is often used and is a moderately conservative. s is often derived from literature review or is assumed conservatively.

$$DE = 1 + ICC \times (s - 1) \quad (13)$$

$$N_{Dc} = \frac{N_D \times DE}{s} \quad (14)$$

4. Using MRMCsampleSize

The function within the **MRMCsampleSize** package that can be used for estimating sample sizes for fully-crossed MRMC studies is `sampleSize_MRMC`. Details of the arguments of this function are detailed in Table 2.

4.1. Example 1: Without adjustment for ICC

Scenario: An AI software to detect and localize adenomas from CT colonography scans is available. Will using this device increase the accuracy (in terms of AUC) of radiologists (readers) in detecting adenomas in CT colonography? Only case-level AUCs are to be estimated and correlation arising from the presence of adenomas in a single scan can be ignored. Determine a sample size for an MRMC study considering readers with varying experience (general radiologists, radiologists with abdominal radiology fellowships, 1-12 yrs experience)

Sample size estimation result: A sample of 20 readers and 436 scans (218 with presence of adenomas, 218 with no adenomas) will have 80% power at a type I error rate of 5% to detect a minimum difference in readers' AUC of 5% assuming a large inter-reader and intra-reader variability

Table 2: Function arguments and definitions

Argument	Definition
<i>endpoint</i>	Character string to inform what is the endpoint of the MRMC study. Values can be either <i>auc</i> or <i>sensitivity</i>
<i>J</i>	The number of readers for the study. It is recommended to have minimum 5 readers in any MRMC study.
<i>delta</i>	Effect size denoting the anticipated difference in the endpoint between the two interventions/imaging-modalities/techniques. Typically chosen values are 0.04, 0.05 and 0.06. Should be between 0 and 1.
<i>rangeb</i>	Inter-reader variability (sometimes referred to as between-reader variability) range denoting the anticipated difference between the highest accuracy of any reader in the study and the lowest accuracy of any reader in the study. Should be a numeric value between 0 and 1).
<i>rangew</i>	Intra-reader variability range (sometimes referred to as within-reader variability) denoting the anticipated difference between the accuracies of a reader who interprets the same images using the same imaging technique at two different times. Should be a numeric value between 0 and 1.
<i>theta</i>	Expected average value of the endpoint for the <i>J</i> readers. Should be a numeric value between 0 and 1.
<i>R</i>	Ratio of non-diseased cases to diseased cases. Defaults to 1.
<i>r1</i>	Correlation between FOMs of readers when same cases are evaluated by the same reader using different modalities.
<i>r2</i>	Correlation between FOMs when the same cases are evaluated by different readers using the same modality. It is assumed that $r2 = r3$ for default calculations
<i>r3</i>	Correlation between FOMs when the same cases are evaluated by different readers using different modalities. It is assumed that $r2 = r3$ for default calculations.
<i>rb</i>	Correlation between FOMs when the same readers evaluate cases using different modalities. Defaults to 0.8.
<i>K</i>	Number of times each reader interprets the same images from the same modality. $K = 1$ in a fully-crossed paired-reader paired-case study design with two modalities.
<i>power</i>	Power to detect <i>delta</i> given all other assumptions. Default value is 0.8 corresponding to 80 percent power.
<i>alpha</i>	The type I error rate. Default value is 0.05 corresponding to 5 percent type I error (significance level).
<i>nu1</i>	Numerator degrees of freedom of the F-distribution which will be used to estimate the non-centrality parameter. Defaults to 1.
<i>var_auc</i>	Variance estimation method when endpoint is <i>auc</i> . Defaults to <i>obuchowski</i> based on Zhou et.al. (2011). If value is changed to <i>blume</i> , then method proposed by Blume (2009) will be used to estimate the variance.
<i>reader_var_estimation_method</i>	Method to be used to estimate inter- and intra-reader variances from <i>rangeb</i> and <i>rangew</i> . Defaults to <i>normal</i> which corresponds to the first method in section 3.2.2.1.
<i>n_reading_sessions_per_reader</i>	Number of times each reader interprets each case. Defaults to 2 which corresponds to a typical MRMC study with 2 modalities.
<i>corr</i>	Logical value indicating if ICC has to be adjusted or not. Defaults to <i>FALSE</i> which indicates adjustment is not required. If <i>TRUE</i> , then both <i>ICC</i> and <i>s</i> have to be specified as valid numerical values.
<i>ICC</i>	A numerical value between 0 and 1 indicating the expected ICC.
<i>s</i>	Average number of lesions in diseased cases. Should be a numeric value which is more than 1.

of 20% and 5% respectively, a 0.47 moderate correlation between readers, anticipated average readers' AUC as 0.75.

The programmatic implementation of sample size estimation for the example is shown below:

```
ex1 <- sampleSize_MRMC(endpoint = 'auc',
                       J = 20,
                       delta = 0.05,
                       theta = 0.75,
                       r1 = 0.47,
                       rangeb = 0.20,
                       rangew = 0.05)
ex1$ORSampleSizeResults

#>
#>      Obuchowski-Rockette Sample Size Estimation Results
#>
#>      ICC = Not applicable
#>      nUnits_i = 218
#>      nCases_c = NA
#>      nControls = 218
#>      nTotal = 436
#>      J = 20
#>      DE = NA
#>      s = NA
#>      power = 0.8
#>      alpha = 0.05
#>
#> NOTE:
#> ICC: Is intra-cluster correlation (ICC) considered while estimating sample size?
#> nUnits_i: Number of required units (a unit is a lesion of interest) assuming independence between units
#> nCases_c: Number of required diseased cases with presence of at least one unit of lesion after adjusting for ICC
#> nControls: Number of required non-diseased cases
#> nTotal: Total sample size (cases)
#> J: Number of readers
#> DE: Design effect due to ICC
#> s: Assumed average number of lesions in diseased cases
#> power: Assumed power
#> alpha: Significance level (Type I error rate)
```

4.2. Example 2: With adjustment for ICC

Scenario: A single CT colonography scan can have multiple adenomas. The scenario is same as in example 1, but with adjustment for correlation is needed. Not only that the radiologist need to detect a lesion (s), they also have to localize the detected lesion (s) correctly. The task is to determine a sample size for an MRMC study considering readers with varying experience (general radiologists, radiologists with abdominal radiology fellowships, 1-12 yrs experience) with primary endpoint being difference in AUC of readers when aided by the AI device as compared to when not aided by it.

Sample size estimation result: A sample of 20 readers and 394 scans (197 with presence of at least one adenoma, 197 with no adenoma) will have 80% power at a type I error rate of 5% to detect a minimum difference in readers' AUC of 5% assuming a large inter-reader and intra-reader variability of 20% and 5% respectively, a 0.47 moderate correlation between readers, anticipated average readers' AUC as 0.75, a moderate ICC of 0.5 due to the presence of multiple nodules in a single scan and average number of adenomas in scans with adenomas as 1.25.

```
ex2 <- sampleSize_MRMC(endpoint = 'auc',
                       J = 20,
                       delta = 0.05,
                       theta = 0.75,
                       r1 = 0.47,
                       rangeb = 0.20,
                       rangew = 0.05,
                       corr = TRUE,
                       ICC = 0.5,
```



```
      s = 1.25)
print(ex2$ORSampleSizeResults)

#>
#>      Obuchowski-Rockette Sample Size Estimation Results
#>
#>      ICC = Intra-class correlation applicable
#>      nUnits_i = 218
#>      nCases_c = 197
#>      nControls = 197
#>      nTotal = 394
#>      J = 20
#>      DE = 1.125
#>      s = 1.25
#>      power = 0.8
#>      alpha = 0.05
#>
#> NOTE:
#> ICC: Is intra-cluster correlation (ICC) considered while estimating sample size?
#> nUnits_i: Number of required units (a unit is a lesion of interest) assuming independence between units
#> nCases_c: Number of required diseased cases with presence of at least one unit of lesion after adjusting for ICC
#> nControls: Number of required non-diseased cases
#> nTotal: Total sample size (cases)
#> J: Number of readers
#> DE: Design effect due to ICC
#> s: Assumed average number of lesions in diseased cases
#> power: Assumed power
#> alpha: Significance level (Type I error rate)
```

5. Validation

In order to validate the results from [MRMCsamplesize](#), the sample size numbers from published literature (Nancy A. Obuchowski 2000) were used. The results from [MRMCsamplesize](#) was compared against this published literature. A total of 162 MRMC design scenarios were simulated by varying the number of readers, inter- and intra-reader variability ranges, effect size (delta), ratio of non-diseased to diseased cases and anticipated average AUC by assuming independence (without any adjustment for ICC). The results from the R function were found to be comparable to the numbers from a published reference. Of the 162 simulations, 11 sample size outputs from R (7%) were moderately different from that of the reference value (overestimated by >5% or underestimated by more than 1%). The maximum absolute percentage deviation was 5.26% from the reference number. Most of the differences were found in the small number of reader scenarios (J=4) when assumed reader variabilities are large. Such a scenario is usually not advised anyway as a minimum of five readers is typically needed to get a reliable estimate of the variances in the accuracy of readers. The tables listing the results from this validation exercise are available in the supplementary material (S1).

6. Acknowledgements

The authors want to acknowledge and thank the colleagues at Qure.ai for their support and review of the manuscript. Many existing reference literature provided the much-needed inputs to this work which would have been impossible in the absence of such literature. In that regard, we want to especially acknowledge the authors of the original Obuchowski-Rockette model.

References

- Blume, Jeffrey D. 2009. "Bounding Sample Size Projections for the Area Under a ROC Curve." *Journal of Statistical Planning and Inference* 139 (3): 711–21. <https://doi.org/https://doi.org/10.1016/j.jspi.2007.09.015>.
- CDRH. 2022. "Clinical Performance Assessment CADe 510(k) Submissions Guidance." <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-performance-assessment-considerations-computer-assisted-detection-devices-applied-radiology>.
- Chakraborty, Dev, and Xuetong Zhai. 2023. *RJafroc: Artificial Intelligence Systems and Observer Performance*. <https://dpc10ster.github.io/RJafroc/>.

- Dorfman, D. D., K. S. Berbaum, and C. E. Metz. 1992. "Receiver operating characteristic rating analysis. Generalization to the population of readers and patients with the jackknife method." *Investigative radiology* 27 (9): 723–31.
- FDA. 2023. "AI/ML Enabled Medical Devices." <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices>.
- Gallas, Brandon D. 2023. "iMRMC: Multi-Reader, Multi-Case Analysis Methods (ROC, Agreement, and Other Metrics)." <https://cran.r-project.org/web/packages/iMRMC/index.html>.
- Gallas, Brandon D., Gene A. Pennello, and Kyle J. Myers. 2007. "Multireader Multicase Variance Analysis for Binary Data." *Journal of the Optical Society of America. A, Optics, Image Science, and Vision* 24 (12): B70–80. <https://doi.org/10.1364/josaa.24.000b70>.
- Hillis, Stephen L., and Kevin M. Schartz. 2018. "Multireader Sample Size Program for Diagnostic Studies: Demonstration and Methodology." *Journal of Medical Imaging* 5 (04): 1. <https://doi.org/10.1117/1.jmi.5.4.045503>.
- Iowa, University of. 2023. "Introduction to the Obuchowski-Rockette (OR) and Dorfman-Berbaum-Metz (DBM) Methods of Analysis for MRMC Data." <https://perception.lab.uiowa.edu/introduction-obuchowski-rockette-or-and-dorfman-berbaum-metz-dbm-methods-analysis-multireader>.
- Langlotz, Curtis P. 2019. "Will Artificial Intelligence Replace Radiologists?" *Radiology: Artificial Intelligence* 1 (3): e190058. <https://doi.org/10.1148/ryai.2019190058>.
- Obuchowski, Nancy A. 2000. "Sample Size Tables for Receiver Operating Characteristic Studies." *AJR* 175 (3). <https://doi.org/10.2214/ajr.175.3.1750603>.
- Obuchowski, Nancy A., and Jennifer Bullen. 2022. "Multireader Diagnostic Accuracy Imaging Studies: Fundamentals of Design and Analysis." *Radiology* 303 (1): 26–34. <https://doi.org/10.1148/radiol.211593>.
- Obuchowski, Nancy A., and Stephen L. Hillis. 2011. "Sample Size Tables for Computer-Aided Detection Studies." *American Journal of Roentgenology* 197 (5): 821–28. <https://doi.org/10.2214/AJR.11.6764>.
- Obuchowski, Nancy A., and Howard E. Rockette. 1995. "Hypothesis Testing of Diagnostic Accuracy for Multiple Readers and Multiple Tests: An Anova Approach with Dependent Observations." *Communications in Statistics - Simulation and Computation* 24 (2): 285–308. <https://doi.org/10.1080/03610919508813243>.
- Robert, Dennis. 2023. "MRMCsamplesize: An r Package for Estimating Sample Sizes for Multi-Reader Multi-Case Studies." <https://cran.r-project.org/web/packages/MRMCsamplesize/index.html>.
- Smith, Brian J., and Stephen L. Hillis. 2020. "Multi-Reader Multi-Case Analysis of Variance Software for Diagnostic Performance Comparison of Imaging Modalities." In, edited by Frank Samuelson and Sian Taylor-Phillips, 113160K. <https://doi.org/10.1117/12.2549075>.
- Smith, Brian J., Stephen L. Hillis, and Lorenzo L. Pesce. 2023. "MRMCAov: R Package for Multi-Reader Multi-Case Analysis of Variance." <https://cran.r-project.org/web/packages/MRMCaov/index.html>.
- Zhou, Xiao Hua, Nancy A. Obuchowski, and Donna K. McClish. 2011. "Statistical Methods in Diagnostic Medicine." *Statistical Methods in Diagnostic Medicine*, April, 1–545. <https://doi.org/10.1002/9780470906514>.

Dennis Robert
Qure.ai
Clinical Research Scientist
Bangalore, India
ORCID: 0000-0002-8539-6408
dennis.robert.nm@gmail.com

Saigopal Sathyamurthy
Qure.ai
Clinical Research Scientist
Bangalore, India
saigopal.s@qure.ai

Preetham Putha
Qure.ai
Chief AI Officer
Bangalore, India
preetham.putha@qure.ai